Environmental Health Perspectives
Vol. 32, pp. 169-179, 1979

Some Applications of Categorical Data Analysis to Epidemiological Studies

by James E. Grizzle* and Gary G. Koch*

Several examples of categorized data from epidemiological studies are analyzed to illustrate that more informative analysis than tests of independence can be performed by fitting models. All of the analyses fit into a unified conceptual framework that can be performed by weighted least squares. The methods presented show how to calculate point estimate of parameters, asymptotic variances, and asymptotically valid χ^2 tests. The examples presented are analysis of relative risks estimated from several 2 × 2 tables, analysis of selected features of life tables, construction of synthetic life tables from cross-sectional studies, and analysis of dose-response curves.

Introduction

During the past 25 years there has been a remarkable resurgence in the development of statistical methods for the analysis of categorized data. The methods available are comparable in flexibility and analytical power to those commonly used for intervally scaled data.

It is a measure of the development in this area that well written reasonably comprehensive text books are available; for example Bishop, Fienberg, and Holland (*1*), and Fienberg (*2*). Powerful computer programs for a variety of analyses are also available from several sources. Methods of analysis based on weighted least squares, maximum likelihood and minimum discrimination information are being vigorously pursued.

These methods may yield different analytical results in small samples even though they are asymptotically equivalent. However, the small sample properties of these approaches still represents an area of research where not much definitive is known.

In the authors' opinion, the importance of these developments lie in their ability to go beyond the usual tests of association by permitting the examination of more specific hypotheses about single or several multiway tables. Their generality allows data of different types for which the conventional hypotheses of association or homogeneity are of little interest to be analyzed in ways that are more relevant.

In most cases estimation and testing can be performed by weighted least squares (WLS), maximum likelihood (ML), or minimum discrimination information. In some instances WLS and ML can be combined in useful ways.

In this paper we shall present some examples of the application of WLS methods to data arising in epidemiological investigations.

Notation

To fix ideas, consider the hypothetical data shown in Table 1 and the expected cell probabilities shown in Table 2. The development which follows is based on the methodology discussed in more detail in Grizzle, Starmer, and Koch (*3*), which hereafter is abbreviated as GSK.

Table 1. Frequency distribution.

Populations (factors)	Categories of response				
	1	2	. . .	r	Total
1	n_{11}	n_{12}	. . .	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	. . .	n_{2r}	$n_{2.}$
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
s	n_{s1}	n_{s2}	. . .	n_{sr}	$n_{s.}$

*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514.

Table 2. Expected cell probabilities.

Populations (factors)	Categories of response				
	1	2	...	r	Total
1	π_{11}	π_{12}	...	π_{1r}	1
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
s	π_{s1}	π_{s2}		π_{sr}	1

We define

$$\underset{1\times r}{\boldsymbol{\pi}_i'} = [\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir}]$$

$$\underset{1\times rs}{\boldsymbol{\pi}} = [\boldsymbol{\pi}_1', \boldsymbol{\pi}_2', \ldots, \boldsymbol{\pi}_s']$$

and their sample proportion analogs as

$$\mathbf{p}_{ij} = n_{ij}/n_{i.}$$

$$\underset{1\times r}{\mathbf{p}_i'} = [p_{i1}, p_{i2}, \ldots, p_{ir}]$$

$$\underset{1\times rs}{\mathbf{p}'} = [\mathbf{p}_1', \mathbf{p}_2', \ldots, \mathbf{p}_s']$$

Let

$$\text{var}(p_i) = \underset{r\times r}{\mathbf{V}}(\boldsymbol{\pi}_i) = \frac{1}{n_{1.}}$$

$$\begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \ldots & -\pi_{i1}\pi_{ir} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \ldots & -\pi_{i2}\pi_{ir} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ -\pi_{ir}\pi_{i1} & -\pi_{ir}\pi_{i2} & \ldots & \pi_{ir}(1-\pi_{ir}) \end{bmatrix}$$

where

$\underset{r\times r}{\mathbf{V}}(\mathbf{p}_i)$ = sample estimate of $\mathbf{V}(\boldsymbol{\pi}_i)$

$\underset{rs\times rs}{\mathbf{V}}(\mathbf{p})$ = block diagonal matrix having $\mathbf{V}(\mathbf{p}_i)$ on the main diagonal

$f_m(\boldsymbol{\pi})$ = any function of the elements of $\boldsymbol{\pi}$ that has partial derivatives up to second order with respect to the π_{ij}, $m = 1, 2, \ldots, u \leq (r-1)s$;

$f_m(\mathbf{p}) = f_m(\boldsymbol{\pi})$ evaluated at $\boldsymbol{\pi} = \mathbf{p}$

$[\mathbf{F}(\boldsymbol{\pi})]' = [f_1(\boldsymbol{\pi}), f_2(\boldsymbol{\pi}), \ldots, f_u(\boldsymbol{\pi})]$

$\mathbf{F}' = [F[(\mathbf{p})]' = [f_1(\mathbf{p}), f_2(\mathbf{p}), \ldots, f_u(\mathbf{p})]$;

$$\underset{u\times rs}{\mathbf{H}} = \left[\left.\frac{\partial f_m(\boldsymbol{\pi})}{\partial \pi_{ij}}\right|_{\pi_{ij}=p_{ij}}\right]$$

$$\underset{u\times u}{\mathbf{S}} = \mathbf{HV}(\mathbf{p})\mathbf{H}'$$

We assume that the functions $f_i(\boldsymbol{\pi})$ are jointly independent of one another and of the constraint

$$\sum_{j=1}^{r} \pi_{ij} = 1 \qquad i = 1, 2, \ldots, s$$

i.e., both $\mathbf{H}$ and $\mathbf{HV}(\boldsymbol{\pi})\mathbf{H}'$ are of rank u. When these conditions hold, then $\mathbf{S}$ is expected to have rank u. However, for some types of data, if some of the $n_{ij} = 0$, S will be of rank less than u. Therefore, if difficulty is created by an occasional $n_{ij} = 0$, we follow Berkson (*4*) and suggest that it be replaced by $1/r$. This has the effect of making the estimate of π_{ij} be $1/rn_{i.}$, which is the extension of Berkson's procedure to the multinomial case. However, we have made no extensive investigation of the effect of this rule in the multinomial case such as Berkson did for the binomial case. Alternatively, the combined WLS and ML methods described in Koch et al. (*5*) can be used to bypass this type of difficulty (see Example 4).

Estimation and Testing

We assume that

$$\underset{u\times 1}{\mathbf{F}(\boldsymbol{\pi})} = \underset{u\times v}{\mathbf{X}}\ \underset{v\times 1}{\boldsymbol{\beta}} \qquad (1)$$

where $\mathbf{X}$ is a known design matrix (which may be different from the usual design matrix in the sense of reflecting a multivariate framework when more than one function is constructed within each population as will be illustrated later) of rank $v \leq u$ and $\boldsymbol{\beta}$ is a vector of unknown parameters.

Several workers have shown that if the hypothesized model fits the data, a best asymptotic normal (BAN) estimate of $\boldsymbol{\beta}$ is given by $\mathbf{b}$, when $\mathbf{b}$ is the vector which minimizes $(\mathbf{F} - \mathbf{Xb})'\,\mathbf{S}^{-1}\,(\mathbf{F} - \mathbf{Xb})$. The minimum value of this form may be used to test the fit of the model $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$. Given that the presumed model provides an adequate fit to the data, a test of the hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{O}$ is produced by conventional methods of weighted multiple regression, where $\mathbf{C}$ is a $(d \times v)$ matrix of arbitrary constants of full rank $d \leq v$.

The test statistic for the fit of the model is

$$\text{SS}[\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}] = \mathbf{F}'\mathbf{S}^{-1}\,\mathbf{F} - \mathbf{b}'\,(\mathbf{X}'\mathbf{S}^{-1}\,\mathbf{X})\mathbf{b} \quad (2)$$

which has asymptotically a (central) χ^2-distribution with D.F. $= (u - v)$ if the model fits, where $\mathbf{b} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{S}^{-1}\mathbf{F}$. Given the model, the test of the hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is produced by

$$\text{SS}[\mathbf{C}\boldsymbol{\beta} = \mathbf{0}] = \mathbf{b}'\mathbf{C}'\,[\mathbf{C}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{Cb}$$

which has asymptotically a χ^2-distribution with D.F. $= d$ if H_0 is true.

In many cases there is only one population, and the objective of the statistical analysis is to study the relationships among several ways of classification of

the sample units. Many tests appropriate to this problem can be formulated as

$$\underset{u \times 1}{\mathbf{F}}\ (\pi) = \mathbf{0}.$$

This fits into the general framework by setting $\mathbf{X} = \mathbf{I}_u$, the identity matrix so that $\mathbf{b} = \mathbf{F}$ and using $\mathbf{C} = \mathbf{I}_u$ so that the test statistic is $\mathbf{F}'\mathbf{S}^{-1}\mathbf{F}$, which has asymptotically a χ^2-distribution with D. F. $= u$ if H_0 is true.

Special Cases of $f(\pi)$

The form of S depends on H and through H on the function $\mathbf{F}(\pi)$. Therefore for each family of functions $\mathbf{F}(\pi)$, S will be different. For linear relationships, one can define a family of functions

$$\underset{u \times rs}{\mathbf{A}}\ \underset{rs \times 1}{\pi}$$

where A [Eq. (3)]

$$\mathbf{A} = \begin{bmatrix} a_{111}\ a_{112}\ \cdots\ a_{11r};\ a_{121}\ a_{122}\ \cdots\ a_{12r};\ \ldots;\ a_{1s1}\ a_{1s2}\ \cdots\ a_{1sr} \\ a_{211}\ a_{212}\ \cdots\ a_{21r};\ a_{221}\ a_{222}\ \cdots\ a_{22r};\ \ldots;\ a_{2s1}\ a_{2s2}\ \cdots\ a_{2sr} \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ a_{u11}\ a_{u12}\ \cdots\ a_{u1r};\ a_{u21}\ a_{u22}\ \cdots\ a_{u2r};\ \ldots;\ a_{us1}\ a_{us2}\ \cdots\ a_{usr} \end{bmatrix} \tag{3}$$

is of rank $u \leq s(r-1)$; the a_{rij} are arbitrary constants. For logarithmic relationships, one can define the family of functions

$$\underset{t \times 1}{\mathbf{F}}\ (\pi) = \underset{t \times u}{\mathbf{K1}}\ \log\ \underset{u \times rs}{\mathbf{A}}\ \underset{rs \times 1}{\pi} \tag{4}$$

the α-th element of $F(\pi)$ has the form

$$F_\alpha(\pi) = \sum_{\gamma=1}^{u} k_{\alpha\gamma} \log \left(\sum_{i,j} a_{\gamma ij}\pi_{ij} \right) \tag{5}$$

where the $a_{\gamma ij}$ and $k_{\alpha\gamma}$ are the appropriate elements of A and K, respectively. Here, K is a matrix of arbitrary constants of rank $t \leq j \leq rs$. Some care must be exercised to make sure that the H associated with the functions described above is of full rank (i.e., of rank u for the linear case and of rank t for the logarithmic case).

The matrix of partials of the first transformation $\mathbf{F}(\pi) = \mathbf{A}\pi$ is $\mathbf{H} = \partial\mathbf{F}/\partial\pi = \mathbf{A}$, and $\mathbf{S} = \mathbf{A}\hat{\mathbf{V}}(\mathbf{p})\mathbf{A}'$. In the second case $\mathbf{H} = [\partial\mathbf{F}/\partial\pi \mid \pi = \mathbf{p}] = \mathbf{KD}^{-1}\mathbf{A}$ and $\mathbf{S} = \mathbf{KD}^{-1}\ \mathbf{AV}(\mathbf{p})\ \mathbf{A}'\mathbf{D}^{-1}\mathbf{K}'$, where A is as defined previously; and

$$\mathbf{D} = \begin{bmatrix} \mathbf{a}_1'\mathbf{p} & 0 & \cdots & 0 \\ 0 & \mathbf{a}_2'\mathbf{p} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \mathbf{a}_u'\mathbf{p} \end{bmatrix} \tag{6}$$

where $\mathbf{a}_\gamma'$ represents the γ-th row of A.

Examples

Example 1: Examination of Several Multinomial Distributions

Multiway contingency tables often have some of the ways of classification fixed in advance. These tables correspond to designs in which the response to each 'treatment' is represented by a multiway table whose entries have a multinomial distribution. Often the problem is to determine how functions of the multiway table being considered as the response depend on the design variables. An example of data of this type is given in Table 3.

These data were collected in an international study of atherosclerosis (6). The complete table is an example with two dependent variables (responses),

Table 3. Cases of coronary heart disease classified by type of lesion, age, location, and race.

		New Orleans White			Oslo			New Orleans Negro		
Infarct	Age		No	Yes		No	Yes		No	Yes
Myocardial scar	(35-44)	No	9	8	No	7	3	No	4	7
		Yes	6	6	Yes	2	5	Yes	2	3
		e^u = 1.125			e^u = 5.838			e^u = 0.857		
Myocardial scar	(45-54)	No	10	26	No	6	8	No	10	8
		Yes	16	14	Yes	7	11	Yes	14	4
		e^u = 0.337			e^u = 1.179			e^u = 0.357		
Myocardial scar	(55-65)	No	18	47	No	10	22	No	4	13
		Yes	28	21	Yes	39	39	Yes	14	2
		e^u = 0.287			e^u = 0.455			e^u = 0.044		
Myocardial scar	(65-69)	No	3	13	No	5	16	No	0	4
		Yes	11	5	Yes	27	16	Yes	3	2
		e^u = 0.105			e^u = 0.185			e^u = 0.000		

infarct and myocardial scar, which we denote by i and j, respectively, and two independent variables (in the regression sense), age and the combination location and race, denoted k and l. Within each combination of the independent variables we have a 2×2 subtable with observed cell frequencies as shown in Table 4.

Neither of the two marginal totals for this subtable is considered fixed and the corresponding expected cell probabilities are π_{ijkl} where $\Sigma_{i,j}\, \pi_{ijkl} = 1$ for all combinations of k and l. If infarct and scar are independent within this subtable

$$\pi_{11kl}\pi_{22kl}/\pi_{12kl}\pi_{21kl} = 1 \tag{7}$$

or taking the logarithm

$$\Psi_{kl} = \ln \pi_{11kl} - \ln \pi_{12kl} - \ln \pi_{21kl} + \ln \pi_{22kl} = 0 \tag{8}$$

This suggests using u_{kl}, the sample estimate of Ψ_{kl}, as a measure of association. In this problem it would be informative to investigate how u_{kl} depends on age and the combination race-location.

We consider the model

$$E(u_{kl}) = \mu^* + \alpha_k^* + \beta_l^*, \tag{9}$$

where μ^* is the overall mean effect, α_k^*, $k = 1,2,3,4$, is the effect of the k-th age group and β_l^*, $l = 1,2,3$, is the effect of the l-th location-race combination. We reparametrize to incorporate the restrictions $\Sigma\hat{\alpha}_k^* = 0$ and $\Sigma\,\hat{\beta}_l^* = 0$, which is equivalent to calculating the estimates from:

$$E\{\mathbf{u}\} = E\begin{bmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{31} \\ u_{32} \\ u_{33} \\ u_{41} \\ u_{42} \\ u_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & -1 & -1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \mathbf{X}\beta$$

where the elements of $\mathbf{u}$ are the appropriate u_{kl} values taken from Table 3. We substituted (1/4) for the zero n_{1143} in order to avoid a singular $\mathbf{S}$ matrix. The remainder of the analysis is identical to weighted multiple regression. In this case the analysis is particularly simple since $\mathbf{S}$ is a diagonal matrix with $\Sigma_{i,j}$ $(1/n_{ijkl})$ on the diagonal.

The residual from fitting the model is the age × location-race interaction with respect to the measure of association u. In addition, we can investigate how this measure of association depends on age and the location-race combination.

The estimated parameters and the analysis of variance are shown in Tables 5 and 6, respectively.

Given the model the residual sum of squares is distributed as χ^2 with D.F. = 6; thus we conclude that the model fits the data adequately. The various other sums of squares are calculated using the general hypothesis form $\mathbf{C}\xi = 0$. Thus

Table 4. Observed cell frequencies.

		Infarct	
		No	Yes
Myocardial scar	No	n_{11kl}	n_{12kl}
	Yes	n_{21kl}	n_{22kl}
			$n_{..kl}$

Table 5. Estimated parameters and their standard errors for the data in Table 3.

Parameter	Estimate	Standard error
μ	−1.036	0.236
α_1	1.496	0.443
α_2	0.281	0.343
α_3	−0.413	0.296
β_1	−0.021	0.270
β_2	0.776	0.293

Table 6. Analysis of variance for the data in Table 3.

Source of variation	DF	SS	
Age groups	3	16.61	
Linear trend of age	1		16.28
Remainder	2		0.33
Race and location combinations	2	7.06	
N.O. white vs. N.O. Negro	1		1.59
N.O. white vs. Oslo	1		3.55
N.O. white + N.O. Negro vs. 2 Oslo	1		7.00
Residual	6	2.62	

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{11}$$

yields the test of homogeneity of age groups. To test for approximate linearity of the age effects (note that the last age group covers only a 5-year span), we chose the linear contrast $-3\alpha_1^* - \alpha_2^* + \alpha_3^* + 3\alpha_4^* = 0$. Taking into account the restrictions on the estimates, we find the contrast is estimated by $-6\hat{\alpha}_1 - 4\hat{\alpha}_2 - 2\hat{\alpha}_3$. For testing we might equally well choose $3\hat{\alpha}_1 + 2\hat{\alpha}_2 + \hat{\alpha}_3$, which implies that we could use $\mathbf{C} = (0, 3, 2, 1, 0, 0,)$. We can produce other tests similarly.

We conclude from the analysis that there is no age

by location-race interaction and that the measure of association varies linearly with age. The major difference in race and location combination is between Oslo residents and New Orelans residents as shown by the test statistic for the contrast ($\beta^*_1 - 2\beta^*_2 + \beta^*_3 = 0$).

The above method of analysis can be extended readily to contingency tables having subtables of any number of dimensions with or without some of the probabilities being zero because of a *priori* contraints. More than one function of the probabilities of the subtables can be considered dependent variables as in regression and we can examine how they depend on the independent variables.

Example 2: Analysis of Selected Features of Life Table.

This example has been discussed previously by Koch, Johnson, and Tolley (*7*). They show how to set up a life table in contingency table form that is amenable to analysis by WLS methods. Then they proceed to show the analysis of selected features of a life table can be examined in more detail.

The data were originally discussed by Zippin (*8*) in the context of a project carried out by the End Results Group of the National Cancer Institute, to compare two classification schemes for breast cancer—one due to the International Union Against Cancer and the other due to the American Joint Committee on Cancer Staging and End Results Reporting (*9*). The data on a subgroup of 1233 of the original 2039 were analyzed by Cutler and Myers (*10*) as a statistical examination of the classification of the extent of disease in cancer of the breast.

In their analysis, the cases were classified into 18 subgroups corresponding to the following classification:

1. Degree of skin fixation (S)
   a. None (S_0)
   b. Incomplete (S_1)
   c. Complete (S_2)
2. Node status (N)
   a. Clinically negative (N_0)
   b. Palpable (N_1)
3. Tumor size (T)
   a. 2 cm. or less (T_1)
   b. More than 2 cm but less than 4 cm (T_2)
   c. More than 4 cm (T_3)

Table 7 shows the five year survival rates, and their standard errors for each of the 13 groups.

The variation in the five year survival rates with respect to the factors S, N, and T can be investigated by using the WLS methodology in GSK. Let

$$\mathbf{F}(\hat{\pi}) = \mathbf{G} \tag{12}$$

denote the vector of $u = 18$ five-year survival rates. Then linear regression models of the form

$$\mathbf{F}(\pi) = \mathbf{X}\beta \tag{13}$$

can be fitted by weighted least squares where $\mathbf{X}$ is a specified $(18 \times \nu)$ coefficient or design matrix of rank $\nu \leq 18$, β is the corresponding vector of parameters to be estimated, and weighting is with respect to the reciprocals of the appropriate estimated variances. A goodness-of-fit test of the model and tests of hypotheses are obtained by applying the methods outlined above with $\mathbf{F}$ and $\mathbf{S}$ replaced by $\mathbf{G}$ and $\mathbf{V}_G$.

Table 7. Summary results of each of 18 groups based on different models.

Category	No. of cases	5-yr. survival	Est. std. error	Pred. based on X_2	Pred. based on X_3	Std. error of X_3 predicted value	Residual of X_3 predicted value
$S_0N_0T_1$	195	0.88	0.024	0.90	0.89	0.020	−0.01
$S_0N_0T_2$	226	0.77	0.028	0.76	0.76	0.018	0.01
$S_0N_0T_3$	96	0.62	0.050	0.59	0.64	0.032	−0.02
$S_0N_1T_1$	72	0.78	0.049	0.79	0.77	0.030	0.01
$S_0N_1T_2$	89	0.67	0.050	0.64	0.64	0.026	0.03
$S_0N_1T_3$	53	0.49	0.069	0.47	0.51	0.035	−0.02
$S_1N_0T_1$	41	0.95	0.034	0.90	0.95	0.027	0.00
$S_1N_0T_2$	114	0.74	0.042	0.76	0.72	0.017	0.02
$S_1N_0T_3$	78	0.51	0.057	0.59	0.48	0.036	0.03
$S_1N_1T_1$	24	0.63	0.099	0.79	0.59	0.024	0.04
$S_1N_1T_2$	55	0.58	0.066	0.64	0.59	0.024	−0.01
$S_1N_1T_3$	59	0.57	0.065	0.47	0.59	0.024	−0.02
$S_2N_0T_1$	15	0.93	0.069	0.83	0.90	0.035	0.03
$S_2N_0T_2$	30	0.67	0.086	0.68	0.67	0.031	0.00
$S_2N_0T_3$	26	0.38	0.095	0.51	0.43	0.047	−0.05
$S_2N_1T_1$	7	0.71	0.171	0.71	0.67	0.044	0.04
$S_2N_1T_2$	15	0.47	0.129	0.56	0.55	0.035	−0.08
$S_2N_1T_3$	38	0.39	0.079	0.39	0.42	0.037	−0.03

The first model fitted has a matrix of independent variables.

$$\mathbf{X}_1 = \begin{bmatrix}
1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & 1 & -1 & 1 & 1 & -1 & 0 & 2 & 0 & 0 & 2 & -2 & 0 & 2 & 0 & 0 & 2 & -2 \\
1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & 1 & -1 & -1 & -1 & 1 & 0 & 2 & 0 & 0 & 2 & -2 & 0 & -2 & 0 & 0 & -2 & 2 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 \\
1 & 0 & 2 & 1 & 0 & 2 & 1 & -1 & 0 & 2 & 0 & -2 & 1 & -1 & 0 & 2 & 0 & -2 \\
1 & 0 & 2 & 1 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 4 & 0 & 2 & 0 & 0 & 0 & 4 \\
1 & 0 & 2 & 1 & 0 & 2 & -1 & -1 & 0 & -2 & 0 & -2 & -1 & -1 & 0 & -2 & 0 & -2 \\
1 & 0 & 2 & -1 & 0 & -2 & 1 & -1 & 0 & 2 & 0 & -2 & -1 & 1 & 0 & -2 & 0 & 2 \\
1 & 0 & 2 & -1 & 0 & -2 & 0 & 2 & 0 & 0 & 0 & 4 & 0 & -2 & 0 & 0 & 0 & -4 \\
1 & 0 & 2 & -1 & 0 & -2 & -1 & -1 & 0 & -2 & 0 & -2 & 1 & 1 & 0 & 2 & 0 & 2 \\
1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & -1 & 0 & 2 & 0 & 0 & -2 & -2 & 0 & 2 & 0 & 0 & -2 & -2 \\
1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 & 0 & 2 & 0 & 0 & -2 & -2 & 0 & -2 & 0 & 0 & 2 & 2 \\
1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1
\end{bmatrix}$$

This is a saturated model since the number of effects (i.e., columns of $\mathbf{X}_1$) is equal to the dimension of G. The definition of the columns of $\mathbf{X}_1$ is to some exent arbitrary. However the tests shown in Table 8 are unique in that any other definition of the columns of $\mathbf{X}_1$ that preserved the same vector spaces for the respective sets of effects would yield the same sums of squares. The analysis shown in Table 8 leads to conclusions similar to those obtained by Culter and Myers.

Because of the lack of significance of the two and three way interactions, it is reasonable to fit a model which contains main effects only. This model has a matrix of independent variates $\mathbf{X}_2$ which is shown below.

$$\mathbf{X}_2 = \begin{bmatrix}
1 & 1 & -1 & 1 & 1 & -1 \\
1 & 1 & -1 & 1 & 0 & 2 \\
1 & 1 & -1 & 1 & -1 & -1 \\
1 & 1 & -1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 0 & 2 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & 0 & 2 & 1 & 1 & -1 \\
1 & 0 & 2 & 1 & 0 & 2 \\
1 & 0 & 2 & 1 & -1 & -1 \\
1 & 0 & 2 & -1 & 1 & -1 \\
1 & 0 & 2 & -1 & 0 & 2 \\
1 & 0 & 2 & -1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 \\
1 & -1 & -1 & 1 & 0 & 2 \\
1 & -1 & -1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & -1 \\
1 & -1 & -1 & -1 & 0 & 2 \\
1 & -1 & -1 & -1 & -1 & -1
\end{bmatrix}$$

The results of this analysis are displayed in Table 9. As hoped, the residual goodness of fit is not significant. Tests on the parameters corresponding to $\mathbf{X}_2$ indicate significant ($\alpha = 0.01$) main effects for nodes and tumor size. Predicted values based on $\mathbf{X}_2$ appear

Table 8. Analysis of Variance Based on X_1.

Source of variation	D.F.	X^2
Node status	1	12.09
Skin fixation (total): S	2	5.48
$S \times N$	2	.24
Tumor size (total): T	2	48.11
$S \times T$	4	2.53
$N \times T$	2	4.92
$S \times N \times T$	4	6.37

in column 5 of Table 7. Large residuals with absolute values in excess of 0.05 occur for $S_1N_0T_3$, $S_1N_1T_1$, $S_1N_1T_2$, $S_1N_1T_3$, $S_2N_0T_1$, $S_2N_0T_3$, $S_2N_1T_2$. Although the goodness-of-fit test for this model is nonsignificant, its predictive value is not good.

From this point onwards Koch, Johnson, and Tolley take a different approach. They note that the most important sources of variation are the main effects of nodes and tumor size, but that in addition, the main effects of skin fixation and the node × tumor size interaction have sizeable sums of squares. They proceed to examine node status within each degree of skin fixation and tumor size within the $S \times N$ classifications. Ultimately they arrive at the model defined by $\mathbf{X}^3$ where

$$\mathbf{X}_3 = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & -1 & 0 \\
1 & 1 & -1 & 1 & 0 \\
1 & 1 & -1 & 0 & 0 \\
1 & 1 & -1 & -1 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & -1 \\
1 & 0 & -1 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 \\
1 & -1 & 1 & 0 & 1 \\
1 & -1 & 1 & 0 & 0 \\
1 & -1 & 1 & 0 & -1 \\
1 & -1 & -1 & 1 & 0 \\
1 & -1 & -1 & 0 & 0 \\
1 & -1 & -1 & -1 & 0
\end{bmatrix} \qquad (15)$$

The residual sum of squares $SS[G(x) = X\beta] = 2.76$ with DF = 2.76 implies that this model is a good fit.

The predicted values derived on the basis of $\mathbf{X}_3$ are given in Table 7. In addition, the residuals for this

Table 9. Analysis of variance based on X_2.

Source of variation	D.F.	X_2
Skin fixation: S	2	3.76
Node status: N	1	17.89
Tumor Size: T	2	97.38
Overall	5	154.38
Residual	12	16.30

situation are shown on the last column of Table 7 and are all less than 0.05 in absolute value except for the $S_2N_1T_2$ combination. Thus, they conclude that this model provides a reasonably complete explanation of the variation of the five-year survival rates as a function of the factors skin fixation, node status, and tumor size. The standard errors of the predicted values are estimated by the square roots of the diagonal elements of the matrix

$$\mathbf{V}\hat{\mathbf{G}}) = \mathbf{X}[\mathbf{X}'\mathbf{V}_G^{-1}\mathbf{X}]^{-1}\mathbf{X}' \qquad (16)$$

They are shown in Table 10.

Table 10. Analysis of variance based on X_3.

Source of variation	D.F.	X^2
Skin fixation (linear): *S (L)*	1	6.37
Node status: *N*	1	20.48
Tumor size (linear) in S_0N_1 or S_2N_1	1	39.31
Tumor size (linear) in S_1N_0 or S_2N_0	1	75.25
Overall model	4	167.92
Residual	13	2.76

Example 3: Synthetic Life Table

An alternative to following a cohort and observing them until the occurrence of some well defined event is the study of separate samples drawn from non-overlapping subpopulations each of which corresponds to a specific range of values (e.g., age range) for the overall time period of exposure to risk for the occurrence of the vital event of interest. This sampling is presumed to take place cross-sectionally at a single instant in time when subjects are examined for the presence of the event. An example of such data is shown in Table 11.

These data, from Ashford and Sowden (*11*), are from a survey of working coal miners at a representative sample of collieries distributed throughout the United Kingdom. Each subject was classified as to whether he reported the symptoms of breathlessness and wheeze. Then under the assumption of (1) no remission from the event of interest, i.e., once an individual is observed with the event, he is never observed subsequently without it; (2) migration behavior being statistically independent of the event of interest; (3) occurrence rates for the event of interest being constant over time; the data for the period study in Table 11 can be regarded as a synthetic life table. An additional convenient assumption is (4) the survival rates associated with the respective time points throughout the j-th interval are symmetrically distributed with respect to the midpoint of the interval.

When the conditions (1)-(4) are applicable, the synthetic life table associated with a period study can be analyzed by the same probability models that underly cohort studies. A result of these assumptions is that cohort studies and period studies as well as life table analysis and contingency table analysis can be unified into a common methodological framework.

In the analysis that follows, due to Freeman, Freeman and Koch (*12*), the Weibull distribution was fitted to a cross-sectional (period) study. In this context when t represents the time to the occurrence of an event of interest (e.g., a death or the detection of a tumor), then the Weibull cumulative distribution function may be written as:

$$G(t|\mu,\delta, w) = 1 - \exp\{-\mu(t - w)^{\delta}\}$$
$$\mu, \delta \geq 0 \text{ for } t \geq w; \qquad (17)$$

with the interpretation of the parameters, (μ,δ,w) dependent on the type of data being analyzed.

Most applications of Weibull type distributions is in circumstances in which a carcinogen is applied in a relatively uniform and continuous manner (for example, weekly skin paintings), and the variable of

Table 11. Observed frequency of symptoms and marginal proportions.

Age group	Breathless Yes; Wheeze: Yes $\bar{D}_{1j}$	Breathless Yes; Wheeze: No $\bar{D}_{2j}$	Breathless No; Wheeze: Yes $\bar{D}_{3j}$	Breathless No; Wheeze: No $\bar{D}_{4j}$	Breathless Margin P_{1j}	Wheeze Margin P_{2j}	Survive Wheeze; Given Breathless P_{3j}
20-24	9	7	95	1841	0.9918	0.9467	0.4375
25-29	23	9	105	1654	0.9821	0.9285	0.2813
30-34	54	19	177	1863	0.9654	0.8906	0.2603
35-39	121	48	257	2357	0.9393	0.8642	0.2840
40-44	169	54	273	1778	0.9019	0.8056	0.2421
45-49	269	88	324	1712	0.8508	0.7522	0.2465
50-54	404	117	245	1324	0.7507	0.6895	0.2246
55-59	406	152	225	967	0.6811	0.6394	0.2724
60-64	372	106	132	526	0.5792	0.5563	0.2218

interest is the time to appearance of a tumor. Peto, Lee, and Paige (*13*) interpret the Weibull parameters as follows: μ is a rate-determining scale parameter and δ and w characterize the process by which the tumor develops. Accordingly, hypotheses concerning μ are appropriate for examining whether carcinogens have different intensities, while differences among the δ indicate different processes.

From Eq. (17), it follows that

$$\ln \{1 - G(t|\mu, \delta, w)\} = -\mu(t - w)^{\delta} \quad (18)$$

where t refers either to time or age. It is assumed that w has a fixed known value (e.g., $w = 0$) which can be justified. As a result, a linear model involving the parameters ($\ln \mu$) and δ can be obtained by multiplying both sides of Eq. (17) by (-1) and then applying logarithmic transformations a second time which results in

$$\begin{aligned}\theta(t) &= \ln[-\ln\{1 - G(t|\mu, \delta, w)\}] \\ &= \ln \mu + \delta \ln (t - w) \end{aligned} \quad (19)$$

The weighted least-squares methodology in GSK can be used to fit the model (19) to sample estimates of $\theta(t)$ for which consistent estimates of variance can be obtained by methods described in Forthofer and Koch (*14*). To illustrate this method the data in Table 11 are used. These data have also been analyzed in a number of other papers (*15-17*).

Let x denote age as it ranges from birth ($x = 0$) to 100 years in five-year intervals. Let j index these five-year age groups by corresponding to the right endpoint of the age group so that

$$j = \left\{\frac{x}{5}\right\} = 1, 2, 3, \ldots$$

for $x = 5, 10, 15, \ldots$, respectively. Since this is a period study, it is required that the parameter w be used to shift the survival probabilities back to the midpoint of the age interval. Let $\bar{P}_j$ denote the proportion of subjects in the j-th age interval range who have survived (in the sense of not reporting) breathlessness through the instant in time at which the survey was conducted. Thus the fitted Weibull model for surviving breathlessness is

$$\{\bar{P}_j \hat{=} \exp(-\mu(j - 0.5)^{\delta}\} \quad (20)$$

where the symbol $\hat{=}$ means, "is an estimate of." Then

$$F(\bar{P}_j) = \ln \{-\ln\{\bar{P}_j\}\} \hat{=} \ln \mu + \delta \ln (j - 0.5) \quad (21)$$

In matrix notation, let $\mathbf{F} = F(\bar{P}_5), F(\bar{P}_6), \ldots, F(\bar{P}_{13})$ be our vector of functions of estimated survival probabilities. Then it may be written as

$$\mathbf{F} = \mathbf{L}(\ln \{\mathbf{K} \ln [\bar{\mathbf{P}}]\}) \quad (22)$$

where $\bar{\mathbf{P}}' = (\bar{P}_5, \bar{P}_6, \ldots, \bar{P}_{13})$, $\mathbf{K} = -\mathbf{I}_9$, $\mathbf{L} = \mathbf{I}_9$, $\mathbf{I}_9$ is a 9×9 identity matrix, and $\ln_e$ is a natural log function of the vector applied to each component. If one starts with the vector of observed frequencies in Table 11, then $\bar{\mathbf{P}}$ is written, $\bar{\mathbf{P}} = \mathbf{Ap}$, where

$$\mathbf{p}' = (p_{5,1}, p_{5,2}, p_{5,3}, p_{5,4}, p_{6,1}, \ldots, p_{13,1}, p_{13,2}, p_{13,3}, p_{13,4}) \quad (23)$$

and p_{jk} denotes the proportion of subjects in age interval j in symptom class k (k is 1 for breathless and wheeze, 2 for breathless and no wheeze, 3 for no breathless and yes wheeze, and 4 for no symptoms), and $\mathbf{A} = [0\ 0\ 1\ 1] \otimes \mathbf{I}_9$. Here, $\otimes$ denotes Kronecker product. This, $\mathbf{F}$ may then be fitted to the following linear model,

$$E_A\{\mathbf{F}\} = \mathbf{X}\beta = \begin{bmatrix} 1 & \log 4.5 \\ 1 & \log 5.5 \\ \ldots & \ldots \\ 1 & \log 12.5 \end{bmatrix} \begin{matrix} \log \mu \\ \delta \end{matrix} \quad (24)$$

where $E_A\{\ .\ \}$ means asymptotic expectation. The numerical results and fitted values between age 0 and 89 are shown in Table 12. The goodness of fit χ^2-statistic indicates a satisfactory fit and the test of degeneracy of the model (H_0: $\delta = 0$) is highly significant (Table 13).

The original data given by Ashford and Sowden make it possible to consider two symptoms. Using the formulation for the breathlessness margin, it is possible to fit Weibull models to each margin and some appropriate measure of association. Freeman, Freeman, and Koch (*12*) present this method in more detail.

Table 12. Observed and fitted proportions surviving breathlessness.

Age group	Total subjects n_j	Observed surviving breathlessness P_j	Fitted surviving breathlessness $\hat{P}_j$
0-4	—	—	1.0000
5-9	—	—	0.9999
10-14	—	—	0.9994
15-19	—	—	0.9975
20-24	1952	0.9918	0.9928
25-29	1791	0.9821	0.9831
30-34	2113	0.9654	0.9660
35-39	2783	0.9393	0.9385
40-44	2274	0.9019	0.8978
45-49	2393	0.8508	0.8413
50-54	2090	0.7507	0.7678
55-59	1750	0.6811	0.6780
60-64	1136	0.5792	0.5749
65-69	—	—	0.4643
70-74	—	—	0.3539
75-79	—	—	0.2520
80-84	—	—	0.1658
85-89	—	—	0.0996

Table 13.

Parameter	Parameters for breathlessness, fitted model: P. st.	S.E.	X^2	DF
$\ln_{\delta}\mu$	−11.303	0.245		
	4.241	0.105		
Hypothesis test				
Model: $\delta = 0$			1628.022	1
Goodness of fit			6.138	7

Example 4: Dose Response.

The data in Table 14 which appeared in Sugiura and Otake (*18*) show the number of deaths from leukemia (LD) observed at the Atomic Bomb Casualty Commission (ABCC) and the number of individuals who did not die from leukemia (NLD) during 1950-1970 according to age at the time of the atomic bomb and the estimated radiation dosage. The analysis that follows is taken from Landis, Heyman, and Koch (*19*).

The data in Table 14 involve $s = 6$ subpopulations (dose), $r = 2$ response categories (survival status), and $q = 5$ levels of the covariable (age).

For the first step of the analysis the logit model

$$f(\pi_{hi}) = \ln(\pi_{hi1}/\pi_{hi2}) = \mu + \alpha_h + \beta_i \qquad (25)$$

for $h = 1, 2, \ldots, 5$ and $i = 1, 2, \ldots, 6$ is assumed using either the IPF algorithm discussed in Bishop et al. (*1*) or the Newton-Raphson iteration procedure illustrated by Sugiura and Otake (*18*). For these data in Table 14, the maximum likelihood cell estimates under this model are as shown in Table 15. The goodness of fit statistic for model is $Q_L = 27.8$ for the likelihood ratio test or $Q_p = 27.6$ for the Pearson chi-square test.

Alternatively, using the WLS approach outlined in GSK, the goodness-of-fit statistic is $Q_{\mathrm{WLS}} = 22.9$. In each case, these statistics asymptotically follow the chi-square distribution with DF = 20 under the hypothesis of no dose × age interaction. As a result, this hypothesis is accepted at the $\alpha = 0.10$ level of statistical significance. However, it should be noted that many of the expected cell frequencies in Table 15 are rather small (i.e., less than 5), which casts doubt on the strict validity of the goodness-of-fit statistic. We will, however, assume the model with no second order interaction is adequate.

As indicated in Koch et al. (*5*), the maximum likelihood estimator β for the parameters associated with model (25) and its corresponding estimated covariance matrix $\mathbf{V}(\hat{\beta})$ can be determined by applying the WLS computational procedures originally outlined in GSK to the IPF predicted frequencies in Table 15. By writing the MLE proportions corresponding to these cell estimates in vector notation as $\hat{\mathbf{p}} = (\hat{p}_{111}, \hat{p}_{112}, \hat{p}_{121}, \ldots, \hat{p}_{562})$, with the restrictions $\Sigma_j \hat{p}_{hij} = 1$ for $h = 1, 2, \ldots, 5$ and $i = 1, 2, \ldots, 6$, the estimates of the logit functions in Eq. (25) can be expressed in vector form as

$$\mathbf{l} = \mathbf{F}(\hat{\mathbf{p}}) = \mathbf{A}_1 \ln(\hat{\mathbf{p}}) \qquad (26)$$

where $\mathbf{A}_1 = [1 \; -1] \otimes \mathbf{I}_{30}$ and **ln** denotes the natural logarithm of **l** each of the elements of $\hat{\mathbf{p}}$. As a result, the logit model in Eq. (25) can be expressed as

$$\mathbf{E}_A[\mathbf{F}(\hat{\mathbf{p}})] = \mathbf{X}_L\beta, \qquad (27)$$

where $\mathbf{E}_A$ denotes asymptotic expectation and $\hat{\mathbf{B}}$ can be obtained by WLS computations. The design matrix $\mathbf{X}_L$ and its parameter vector **B** which corresponds to the parameterization given by Sugiura and Otake (*18*) can be written (*10*) as:

$$\underset{30\times 10}{\mathbf{X}_L} = \begin{bmatrix} 1_6 & \mathbf{B} & 1_6 & 0_6 & 0_6 & 0_6 \\ 1_6 & \mathbf{B} & 0_6 & 1_6 & 0_6 & 0_6 \\ 1_6 & \mathbf{B} & 0_6 & 0_6 & 1_6 & 0_6 \\ 1_6 & \mathbf{B} & 0_6 & 0_6 & 0_6 & 1_6 \\ 1_6 & \mathbf{B} & -1_6 & -1_6 & +1_6 & -1_6 \end{bmatrix};$$

$$\underset{10\times 1}{\beta} = \begin{bmatrix} \mu \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} \qquad (28)$$

where $\mathbf{B}' = [\mathbf{0}_5, \mathbf{I}_5]$, $\mathbf{1}_r$ denotes a vector of r ones and $\mathbf{0}_r$ denotes a vector of r zeros. The estimated parameters, together with their estimated standard errors, are shown in Table 16 and the ANOVA table is shown in Table 17. These results are identical to those obtained by Sugiura and Otake.

By the two-stage ML/WLS method of Koch et al. (*5*), stepwise model-fitting techniques can be used to arrive at a model which best reflects the variation in the data with a minimum number of parameters. Since age was shown to have a nonsignificant effect it was removed from the model. The new analysis shows that the smoothed probability of leukemia on the logit scale was best characterized by a linear model on the square root of the midpoint on the radiation dosage scale. The final model had a matrix

Table 14. Deaths from leukemia observed at ABCC (1950-1970).

Age	Survival status[a]	Dose, rads: Not in city	0-9	10-49	59-99	100-199	200+
0-9	LD	0	7	3	1	4	11
	NLD	5015	10752	2989	694	418	387
10-19	LD	5	4	6	1	3	6
	NLD	5973	11811	2620	771	792	820
20-24	LD	2	8	3	1	3	9
	NLD	5669	10828	2798	797	596	624
35-49	LD	3	19	4	2	1	10
	NLD	6158	12645	3566	972	694	608
50+	LD	3	7	3	2	2	6
	NLD	3695	9053	2415	655	393	289

[a]LD denotes death from leukemia; NLD denotes nondeath from leukemia.

Table 15. Maximum likelihood cell estimates under logit model assuming no dose × age interaction.

Age	Survival status[a]	Dose, rads: Not in city	0-9	10-49	59-99	100-199	200+
0-9	LD	2.62	9.28	4.11	1.31	2.04	6.63
	NLD	5012.38	10749.71	2987.89	693.69	419.96	391.37
10-19	LD	2.17	7.07	2.50	1.01	2.67	9.59
	NLD	5975.83	11807.93	2623.50	770.99	792.33	816.41
20-34	LD	2.47	7.80	3.22	1.26	2.42	8.83
	NLD	5668.52	10828.20	2797.78	796.74	596.58	624.17
35-49	LD	3.64	12.34	5.55	2.07	3.79	11.61
	NLD	6157.36	12651.66	3564.46	971.93	691.21	606.39
50+	LD	2.10	8.51	3.62	1.35	2.08	5.34
	NLD	3695.90	9051.38	2414.38	655.65	392.92	289.66

[a]LD denotes death from leukemia; NLD denotes nondeath from leukemia.

of independent variables $\mathbf{X}_F$ and parameter vector $\mathbf{B}_F$ which are:

$$\underset{30\times 2}{\mathbf{X}_F} = \begin{bmatrix} \mathbf{1}_6 & \mathbf{S} \\ \mathbf{1}_6 & \mathbf{S} \\ \mathbf{1}_6 & \mathbf{S} \\ \mathbf{1}_6 & \mathbf{S} \\ \mathbf{1}_6 & \mathbf{S} \end{bmatrix};$$

$$\underset{2\times 1}{\mathbf{B}_F} = \begin{bmatrix} \mu \\ \mathrm{B} \end{bmatrix} \qquad (29)$$

where S′ = (0, 2.24, 5.48, 8.66, 12, 25.17.32). The estimates for this model are

$$\hat{\beta}_F = \begin{bmatrix} -7.61 \\ 0.20 \end{bmatrix} \qquad (30)$$

and

$$\mathbf{V}_{\hat{\beta}_F} = \begin{bmatrix} 18.096 & -1.317 \\ -1.317 & 0.160 \end{bmatrix} \times 10^{-3} \qquad (31)$$

The ANOVA table is shown in Table 18. Analogous results would be anticipated for a ML fit by Newton-Raphson or some other direct optimization method.

Finally, the observed and fitted proportions of death from leukemia are shown in Table 19 for each radiation dose subpopulation.

Table 16. Estimated parameters and their estimated standard errors under model X.

Parameter	Estimated parameter	Estimated s.e.
β_2	0.502	0.315
β_3	0.969	0.360
β_4	1.285	0.469
β_5	2.229	0.393
β_6	3.479	0.319
μ	−7.624	0.278
α_1	0.068	0.176
α_2	−0.298	0.179
α_3	−0.113	0.175
α_4	0.190	0.152

Table 17. ANOVA table for model X_L.

Source of variation	DF	FARM WLS test statistic	LR test statistic
Age	4	4.50	4.68
Dose	5	243.44[a]	175.77[a]
Dose × age interaction	20	Not defined[b]	27.83[c]

[a]Means significant at $\alpha \leq 0.01$.
[b]Denotes the log-likelihood ratio statistic obtained from IPF.
[c]Not defined because FARM is applied to no dose × age interaction model predicted frequencies.

Table 18. ANOVA table for model X_F.

Source of variation	D.F	WLS test statistic
Scored dose effect	1	238.56**
Lack of fit for reduction of X_L to X_F	8	6.51

[a]Means significant at $\alpha = 0.01$.

Table 19. Observed and fitted leukemia death rate per 10,000 under model X_F for each radiation dose subgroup.

Radiation dose levels, rad	Leukemic death rate per 10,000		Standard error of predicted rate per 10,000
	Observed	Predicted	
Not in city	4.9	5.0	0.67
0-9	8.2	7.7	0.88
10-49	13.2	14.5	1.33
50-99	18.0	26.9	2.28
100-199	44.7	54.0	5.31
200+	151.6	143.9	20.25

REFERENCES

1. Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. Discrete Multivariate Analysis. M.I.T. Press, Cambridge, Mass., 1975.
2. Fienberg, S. E. The Analysis of Cross-Classified Data. M.I.T. Press, Cambridge, Mass., 1977.
3. Grizzle, J. E., Starmer, C. F., and Koch, G. G. Analysis of categorical data by linear models. Biometrics 25; 498 (1969).
4. Berkson, J. Maximum likelihood and minimum χ^2 estimates of the logistic function. J. Am. Statist. Assoc. 50, 130 (1955).
5. Koch, G. G., Imrey, P. B., Freeman, D. H., Jr., and Tolley, H. D. The asymptotic covariance structure of estimated parameters from contingency table log-linear models. Proceedings of the 9th International Biometric Conference, Boston, Mass., 1976.
6. Strong, J. P., Solber, L. A., and Restrepo, C. Atherosclerosis in persons with coronary heart disease. Lab. Invest. 18: 527 (1968).
7. Koch, G. G., Johnson, W. D., and Tolley, H. D. A linear models approach to the analysis of survival and extent of disease in multidimensional contingency tables. J. Am. Statist. Assoc. 67: 783 (1972).
8. Zippin, C. Comparison of the international and American systems for staging of breast cancer. J. Nat. Cancer Inst. 36: 53 (1966).
9. American Joint Committee on Cancer Staging and End Results Reporting. Clinical Staging System for Cancer of the Breast, 1962.
10. Culter, S. J., and Myers, M. H., Clinical classification of extent of disease in cancer of the breast. J. Nat. Cancer Inst. 37: 193 (1967).
11. Ashford, J. R., and Sowden, R. R. Multi-variate probit analysis. Biometrics 26: 535 (1970).
12. Freeman, D. H., Jr., Freeman, J. L., and Koch, G. G. A modified χ^2 approach for fitting Weibull models to synthetic life tables. Biomet. J. 20: 29 (1978).
13. Peto, R., Lee, P. N. and Paige, W. S. Statistical analysis of continuous carcinogenesis. Brit. J. Cancer 26: 258-61 (1972).
14. Forthofer, R. N. and Koch, G. G. An analysis for compounded functions of categorical data. Biometrics 29: 143 (1973).
15. Grizzle, J. E. Multivariate logit analysis. Biometrics 27: 1057 (1971).
16. Kullback, S., and Fisher, M. Partitioning second-order interaction in three-way contingency tables. J. Roy. Statist. Soc. C22: 172 (1973).
17. Mantel, N., and Brown, C. A logistic reanalysis of Ashford and Sowden's data on respiratory symptoms in British coalminers. Biometrics 29: 649 (1973).
18. Sugiura, N., and Otake, M. An extension of the Mantel-Haenszel Procedure to K 2 × 2 contingency tables and the relation to the logit model. Commun. Statist. 3: 829 (1974).
19. Landis, J. R., Heyman, E. R., and Koch, G. G. Average partial association in three-way contingency tables: a review and discussion of alternative tests. Int. Statist. Rev. 46: 237 (1978).